# Dynamic Regularized Sharpness Aware Minimization in Federated Learning: Approaching Global Consistency and Smooth Landscape


Yan Sun
The University of Sydney
ysun9899@uni.Sydney.edu.au

Li Shen
JD Explore Academy
mathshenli@gmail.com

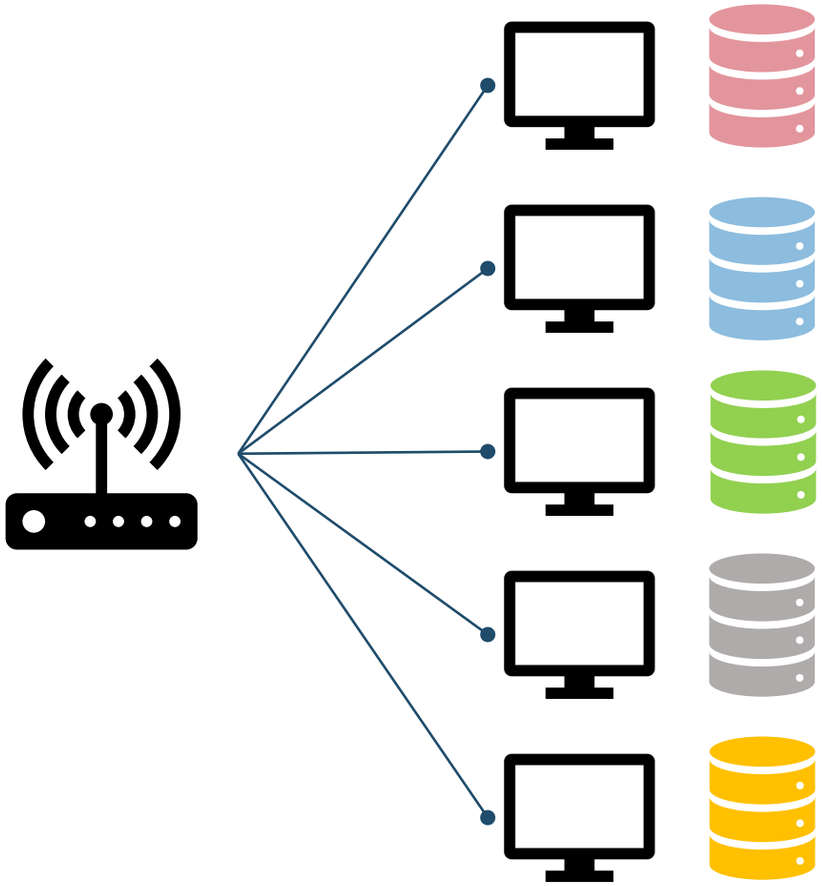
Shixiang Chen
JD Explore Academy
chenshxiang@gmail.com

Liang Ding
JD Explore Academy
liangding.liam@gmail.com

Dacheng Tao
The University of Sydney
dacheng.tao@gmail.com

Section 01 Federated Learning

# Preliminaries – Federated Learning



- Privacy Protection
- Distributed Framework
- Data Shift of Isolated Sources
- Local Limited Calculations
- Communication Bottleneck

## Preliminaries – Problem Setup

$$min \ \ f(w) = \Sigma_i \, p_i f_i(w),$$

$$min \ \ f(w) = \Sigma_i \, p_i f_i(w_i), \quad s.t. \, w_i = w.$$

Where $f_i(w_i) = E_{\{x_i \sim D_i\}} f(w_i; x_i)$ is the ERM loss.

Two main difficulties:

1. Equality constraints : $x_i$ obeys different distribution $D_i$.
2. Resource constraints: limited capacity for computation and communication.

# Preliminaries – FedAvg

**Algorithm 1:** FedAvg Algorithm

**Input:** global model $w$, local model $w_i$, communication round $T$, local interval $K$.
**Output:** The global model $w^T$.

**Initialize states**: initialize $w = w^0$.
**for** $t = 0, 1, ..., T-1$ **do**
  randomly select the active clients set $[n]$ from $[m]$
  **for** $i \in [n]$ ***in parallel*** **do**
    send the $w^t$ to the active clients as $w_{i,0}^t$
    **for** $k = 0, 1, ..., K-1$ **do**
      compute the stochastic gradient $g_{i,k}^t$ at $w_{i,k}^t$
      $w_{i,k+1}^t = w_{i,k}^t - \eta g_{i,k}^t$
    send the $w_i^t = w_{i,K}^t$ to the global server
  $w^{t+1} = \frac{1}{n} \sum_{i \in [n]} w_i^t$
return $w^T$

$[n]$: selected active clients.

$[m]$: total clients.

Section 02 FedSAM Method

# FedSAM Method [1] ---- Preliminaries

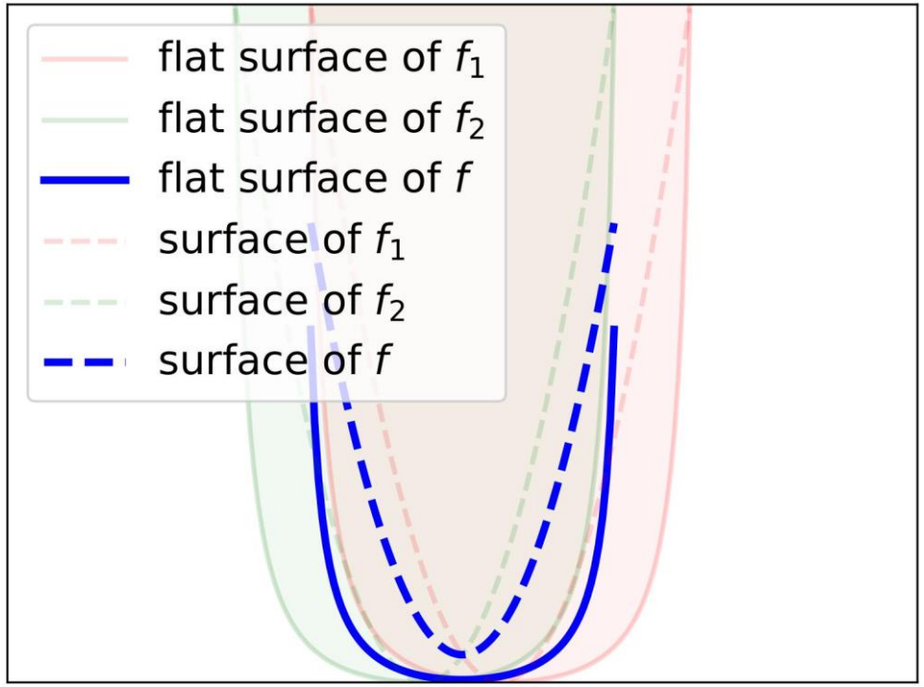
$$min_w \quad f(w) = \sum_{k=0}^{n} p_i f_S(w_i)\,,$$

$$where\ f_S(w_i) = max_{s_i} \quad f_i(w_i + s_i)\ is\ the\ local\ SAM\ objective\ [2]$$

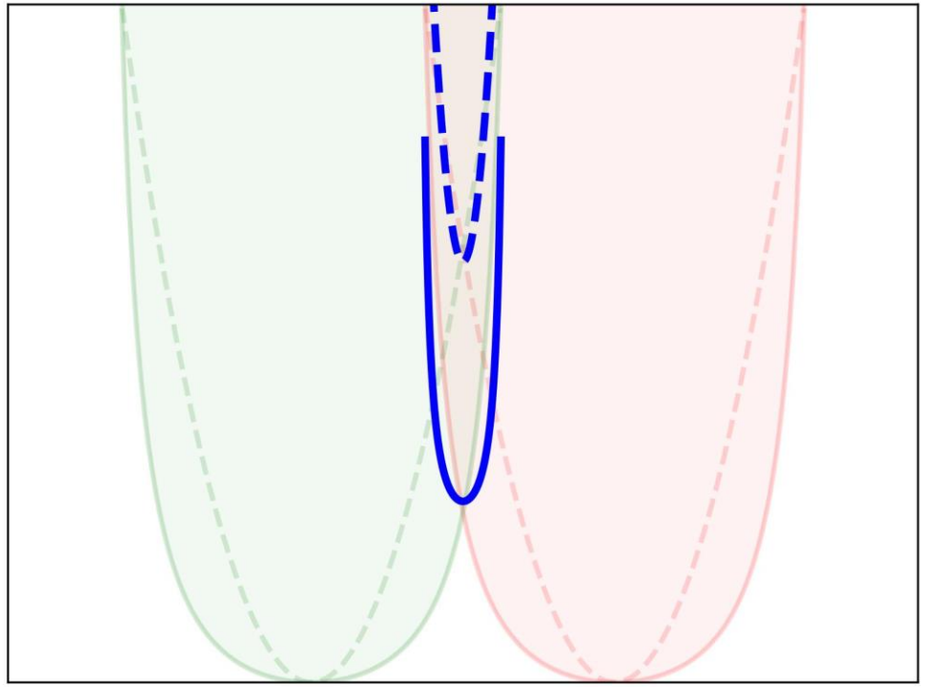
Each local client could achieve a flat optimum with higher generalization performance

[1] Generalized Federated Learning via Sharpness Aware Minimization
[2] Sharpness-Aware Minimization for Efficiently Improving Generalization

# FedSAM Method ---- Challenges




1. Each local optimum is smooth
2. Averaged optimum is smooth



1. Each local optimum is smooth
2. Averaged optimum is STILL SHARP

Section 03

# FedSMOO with Consistent SAM

# FedSMOO Method ---- Preliminaries

$$\min_{\substack{w_i=w \\ s_i=s}} f(w) = \sum_{i=0}^{n} p_i f_S(w_i),$$

*where* $f_S(w_i) = \max_{s_i=s} f_i(w_i + s_i)$ *is the global SAM objective with consistent local optimum* $w_i$ *and perturbation* $s_i$

Each local client could achieve a CONSISTENT flat optimum with higher generalization performance

# FedSMOO Method ---- ADMM Solution

**Lagrangian**

$$\boldsymbol{L}(w, w_i) = \Sigma_i \{ p_i f_S(w_i) + \underbrace{\langle \lambda_i, w_i - w \rangle + \frac{1}{2\beta} \|w_i - w\|^2}_{\text{consistent local models}} + \underbrace{\langle \mu_i, s_i - s \rangle + \frac{1}{2\alpha} \|s_i - s\|^2}_{\text{consistent local perturbation}} \}$$

Double Iterations:

1. Inner: solve the approximation for the maximization in SAM.
2. Outer: solve the minimization

## FedSMOO Method ---- ADMM Solution

### Re-approximate Local SAM

$$s'_i = argmax_{s_i<r} \frac{1}{2\alpha}\|s_i - s'\|^2,\ where\ s' = \alpha(g_{i,k} - \mu_i) - s$$

$$\mu_i = \mu_i + \frac{1}{2\alpha}(s'_i - s')$$

$$s* = argmax_{s<r} \frac{1}{2\alpha}\frac{1}{m}\sum_{i=0}^{n}\|s + \alpha\mu_i - s'_i\|^2$$

### ADMM for Solving FL Objective [3]

$$w_i = argmin_{w_i}\ f_S(w_i) + \langle\lambda_i, w_i - w\rangle + \frac{1}{2\beta}\|w_i - w\|^2$$

$$\lambda_i = \lambda_i - \frac{1}{2\alpha}(w_i - w)$$

$$w = argmin_w \frac{1}{2\beta}\frac{1}{m}\sum_{i=0}^{n}\langle\lambda_i, w_i - w\rangle + \|w - w_i\|^2$$

[3] Federated Learning Based on Dynamic Regularization

# FedSMOO Method ---- Theoretical Analysis

(Convergence) *Let $f$ be the L-smooth non-convex function, $m$ be the number of total clients, $n$ be the number of partial active clients, $r \leq \frac{4\kappa}{\sqrt{nT}}$ where $\kappa$ is a constant, $\beta \leq \frac{\sqrt{n}}{6\sqrt{6m}L}$, the averaged model $w = \sum_{i=0}^{n} w_i$ satisfies:*

$$\frac{1}{T}\sum_{i=0}^{T-1} E||\nabla f(w^t)||^2 \leq \frac{1}{c\beta T}\left[\kappa_f + \frac{1}{n}3\beta L\kappa_r + \frac{m}{n}72\beta^2 L^2\delta^0\right],$$

*where c is a constant in (0,0.5), $\kappa_f = f(w^1) - f^*$, $\delta^0 = \frac{1}{m}\sum_{i=1}^{m} E||w_i - w^1||^2$ is the consistency on the first round.*

(Generalization) *Let $f$ be a non-convex function with $d$ parameters per layer of total $L_l$ layer, the input sample $\varepsilon$ be normalized by the norm of $L_n$ with the size of $D$, using a positive value $p$ and error $e$, with the probability of at least 1-p, the empirical generalization risk on global parameters $w$ is:*

$$G^f \leq \tilde{G}_e^f + O\left(\sqrt{\frac{L_l^2 L_n^2 d\ln(dL_l)V_L + \ln\left(\frac{L_l D}{p}\right)}{(D-1)e^2}}\right),$$

*where $\tilde{G}_e^f$ is the empirical estimate of the above expected margin loss and $V_L = \Pi_{L=1}^{L}\|w_L\|^2 \Sigma_{l=1}^{L}\frac{\|w_l\|_F^2}{\|w_l\|^2}$.*

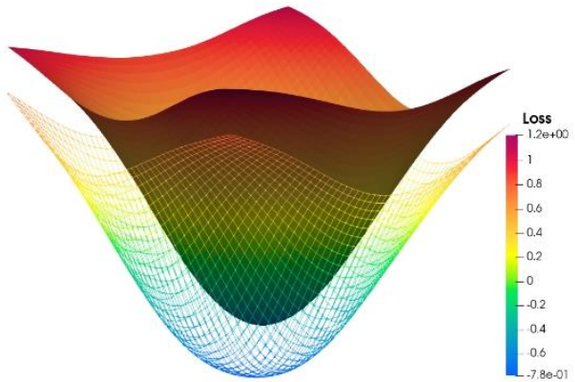
Section 04 Experiments

# FedSMOO – Experiments

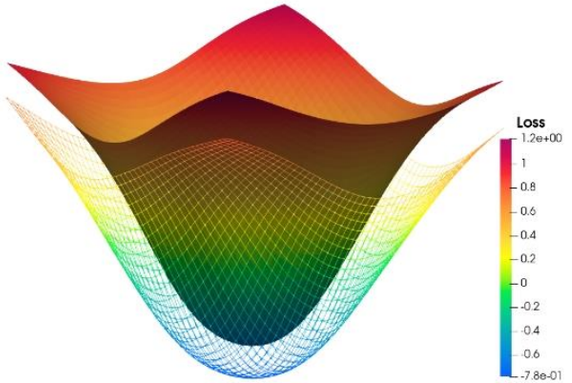
*Table 1.* Test accuracy comparison among baselines and our proposed method on the CIFAR-10/100 dataset after 800 rounds. The dataset splitting method is selected from the Dirichlet sampling with replacement and Pathological partition (only a few random categories are enabled for sampling on a local client). The experimental setups are 10%-100 clients (upper part) and 5%-200 clients (lower part) respectively. "$u$" represents the Dirichlet coefficient which is selected from $[0.1, 0.6]$, and "$c$" represents the number of active categories on each client which is selected from $[3, 6]$ on CIFAR-10 and $[10, 20]$ on CIFAR-100. Each result is calculated by 2 times.

| ALGORITHM | CIFAR-10 | | | | CIFAR-100 | | | |
|---|---|---|---|---|---|---|---|---|
| | DIRICHLET | | PATHOLOGICAL | | DIRICHLET | | PATHOLOGICAL | |
| | $u = 0.6$ | $u = 0.1$ | $c = 6$ | $c = 3$ | $u = 0.6$ | $u = 0.1$ | $c = 20$ | $c = 10$ |
| FEDAVG | $79.52_{\pm.13}$ | $76.00_{\pm.18}$ | $79.91_{\pm.17}$ | $74.08_{\pm.22}$ | $46.35_{\pm.15}$ | $42.64_{\pm.22}$ | $44.15_{\pm.17}$ | $40.23_{\pm.31}$ |
| FEDADAM | $77.08_{\pm.31}$ | $73.41_{\pm.33}$ | $77.05_{\pm.26}$ | $72.44_{\pm.29}$ | $48.35_{\pm.17}$ | $40.77_{\pm.31}$ | $41.26_{\pm.30}$ | $32.58_{\pm.22}$ |
| SCAFFOLD | $81.81_{\pm.17}$ | $78.57_{\pm.14}$ | $83.07_{\pm.10}$ | $77.02_{\pm.18}$ | $51.98_{\pm.23}$ | $44.41_{\pm.15}$ | $46.06_{\pm.22}$ | $41.08_{\pm.24}$ |
| FEDCM | $82.97_{\pm.21}$ | $77.82_{\pm.16}$ | $83.44_{\pm.17}$ | $77.82_{\pm.19}$ | $51.56_{\pm.20}$ | $43.03_{\pm.26}$ | $44.94_{\pm.14}$ | $38.35_{\pm.27}$ |
| FEDDYN | $83.22_{\pm.18}$ | $78.08_{\pm.19}$ | $83.18_{\pm.17}$ | $77.63_{\pm.14}$ | $50.82_{\pm.19}$ | $42.50_{\pm.28}$ | $44.19_{\pm.19}$ | $38.68_{\pm.14}$ |
| FEDSAM | $80.10_{\pm.12}$ | $76.86_{\pm.16}$ | $80.80_{\pm.23}$ | $75.51_{\pm.24}$ | $47.51_{\pm.26}$ | $43.43_{\pm.12}$ | $45.46_{\pm.29}$ | $40.44_{\pm.23}$ |
| MOFEDSAM | $84.13_{\pm.13}$ | $78.71_{\pm.15}$ | $84.92_{\pm.14}$ | $79.57_{\pm.18}$ | $\mathbf{54.38}_{\pm.22}$ | $44.85_{\pm.25}$ | $47.42_{\pm.26}$ | $41.17_{\pm.22}$ |
| **OUR** | $\mathbf{84.55}_{\pm.14}$ | $\mathbf{80.82}_{\pm.17}$ | $\mathbf{85.39}_{\pm.21}$ | $\mathbf{81.58}_{\pm.16}$ | $53.92_{\pm.18}$ | $\mathbf{46.48}_{\pm.13}$ | $\mathbf{48.87}_{\pm.17}$ | $\mathbf{44.10}_{\pm.19}$ |
| FEDAVG | $75.90_{\pm.21}$ | $72.93_{\pm.19}$ | $77.47_{\pm.34}$ | $71.86_{\pm.34}$ | $44.70_{\pm.22}$ | $40.41_{\pm.33}$ | $38.22_{\pm.25}$ | $36.79_{\pm.32}$ |
| FEDADAM | $75.55_{\pm.38}$ | $69.70_{\pm.32}$ | $75.24_{\pm.22}$ | $70.49_{\pm.26}$ | $44.33_{\pm.26}$ | $38.04_{\pm.25}$ | $35.14_{\pm.16}$ | $30.28_{\pm.28}$ |
| SCAFFOLD | $79.00_{\pm.26}$ | $76.15_{\pm.15}$ | $80.69_{\pm.21}$ | $74.05_{\pm.31}$ | $50.70_{\pm.18}$ | $41.83_{\pm.29}$ | $39.63_{\pm.31}$ | $37.98_{\pm.36}$ |
| FEDCM | $80.52_{\pm.29}$ | $77.28_{\pm.22}$ | $81.76_{\pm.24}$ | $76.72_{\pm.25}$ | $50.93_{\pm.31}$ | $42.33_{\pm.19}$ | $42.01_{\pm.17}$ | $38.35_{\pm.24}$ |
| FEDDYN | $80.69_{\pm.23}$ | $76.82_{\pm.17}$ | $82.21_{\pm.18}$ | $74.93_{\pm.22}$ | $47.32_{\pm.18}$ | $41.74_{\pm.21}$ | $41.55_{\pm.18}$ | $38.09_{\pm.27}$ |
| FEDSAM | $76.32_{\pm.16}$ | $73.44_{\pm.14}$ | $78.16_{\pm.27}$ | $72.41_{\pm.29}$ | $45.98_{\pm.27}$ | $40.22_{\pm.27}$ | $38.71_{\pm.23}$ | $36.90_{\pm.29}$ |
| MOFEDSAM | $82.58_{\pm.21}$ | $78.43_{\pm.24}$ | $84.46_{\pm.20}$ | $79.93_{\pm.19}$ | $\mathbf{53.51}_{\pm.25}$ | $42.22_{\pm.23}$ | $42.77_{\pm.27}$ | $39.81_{\pm.21}$ |
| **OUR** | $\mathbf{82.94}_{\pm.19}$ | $\mathbf{79.76}_{\pm.19}$ | $\mathbf{84.82}_{\pm.18}$ | $\mathbf{81.01}_{\pm.19}$ | $53.45_{\pm.19}$ | $\mathbf{45.83}_{\pm.18}$ | $\mathbf{44.70}_{\pm.21}$ | $\mathbf{43.41}_{\pm.22}$ |

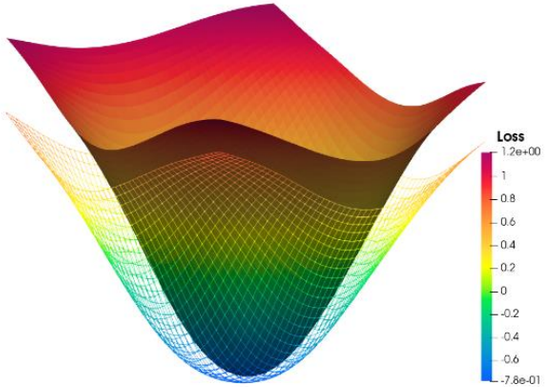
# FedSMOO – Experiments




(a) FedAvg $v.s$ FedSMOO




(b) SCAFFOLD $v.s$ FedSMOO




(c) MoFedSAM $v.s$ FedSMOO

*Table 2.* Consistency and Hessian matrix.

| method | FedSAM | | FedSMOO | |
|---|---|---|---|---|
| Dirichlet | 0.6 | 0.1 | 0.6 | 0.1 |
| $\frac{1}{n}\sum_i \|w_i^t - w^t\|^2$ | 0.866 | 1.245 | 0.821 | 1.061 |
| Hessian Top Eigenvalue | 142.65 | 177.18 | 91.46 | 107.44 |
| Hessian Trace | 3104.1 | 3842.3 | 1783.3 | 2689.4 |

# FedSMOO – Discussions

Main difficulties in the FL:

(1) Local consistency (Equality constraint $w = w_i$).

(2) Local over-fitting on the small dataset (Higher generalization).

(3) Efficient local optimizer design (Generalization-efficiency).

# THANK YOU